Supplementary Material: Quantifying alternative splicing from paired-end RNA-sequencing data

David Rossell [1] **Camille Stephan-Otto Attolini**[1] **Manuel Kroiss**
Almond Stöcker

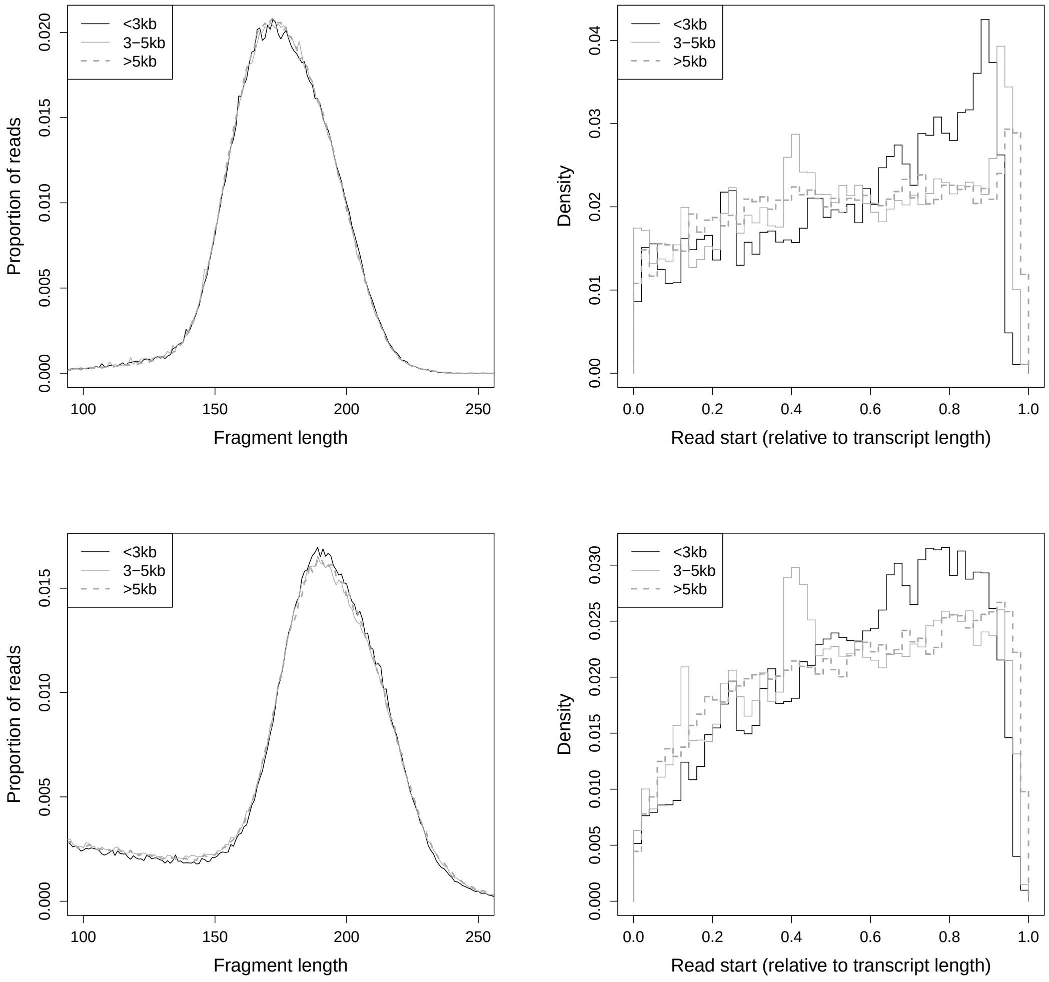
1. Fragment start and length distribution vs. gene length

Two basic elements in our probability model are the distributions of fragment start and length (P_S and P_L, respectively). These determine the probability of observing any given exon path under each splicing variant under consideration, which are the components needed to evaluate the likelihood (main paper, Expression (2)). To our knowledge, all previous methods assumed a common (P_S, P_L) across genes, probably due to the impossibility of obtaining reliable estimates for each individual gene. Following a referee's suggestion, here we assess the validity of that assumption in two experimental data sets. We hypothesized that (P_S, P_L) may depend on the gene length, as it may affect RNA degradation or poly-A tail capture techniques. We split genes into those with length <3,000, 3000-5000 and $>$ 5000 base pairs (bp), and estimated (P_S, P_L) separately for each subset. Once (P_S, P_L) have been estimated, the rest of our approach proceeds as usual. Our R package `casper` includes a function `splitGenomeByLength` to perform this genome splitting operation.

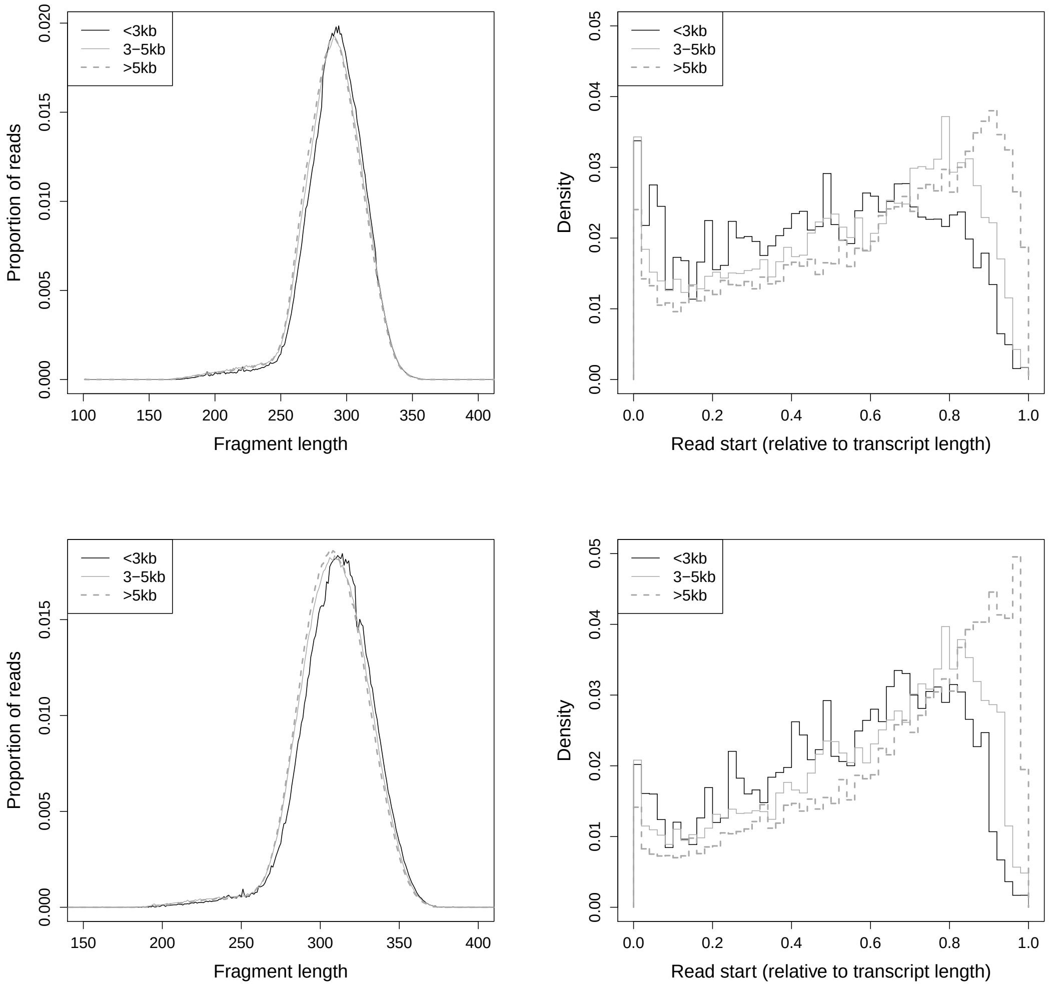
Supplementary Figure 1 shows the estimated distributions for the two RGASP samples (main paper, Section 4.2). The estimated fragment length distribution $\hat{P}_L$ remains virtually identical across the three gene subsets in both samples. However, we observed interesting differences in the start distribution $\hat{P}_S$, shorter genes exhibiting a stronger 3' end bias.

Supplementary Figure 2 shows the estimates for the two Encode samples (main paper, Section 4.3). Here $\hat{P}_L$ remains constant and $\hat{P}_S$ again shows differences according to gene length. Interestingly, we now observe a stronger 3' end bias in longer genes, suggesting that this bias depends on the setup and procedures used in each experiment.

[1]: D.R. and C.S.-O.A. contributed equally to this work



Supplementary Figure 1. *Estimated fragment length (left) and start (right) distribution for two RGASP samples*



Supplementary Figure 2. *Estimated fragment length (left) and start (right) distribution for two Encode samples*

2. MCMC convergence

In order to asses convergence of the proposed MCMC posterior sampling algorithm (Section 3) we compared two independent chains for all genes with reads per kilobase per million (RPKM) above 10 in chromosome 1. We computed the mean absolute difference (MAD) between the posterior mean, 2.5% and 97.5% posterior quantiles for a number of iterations ranging from 10,000 to 100,000 (after a 1,000 burn-in).

Results are shown in Supplementary Figure 3. The MAD between estimated posterior means is below 0.003 for as few as 10,000 iterations, and decreases to roughly 0.001 for 100,000 iterations. Similarly low values are observed for the 2.5% and 97.5% quantiles. Based on these results, a default 10,000 iterations after a 1,000 burn-in may suffice for practical purposes.

3. Further simulation study results

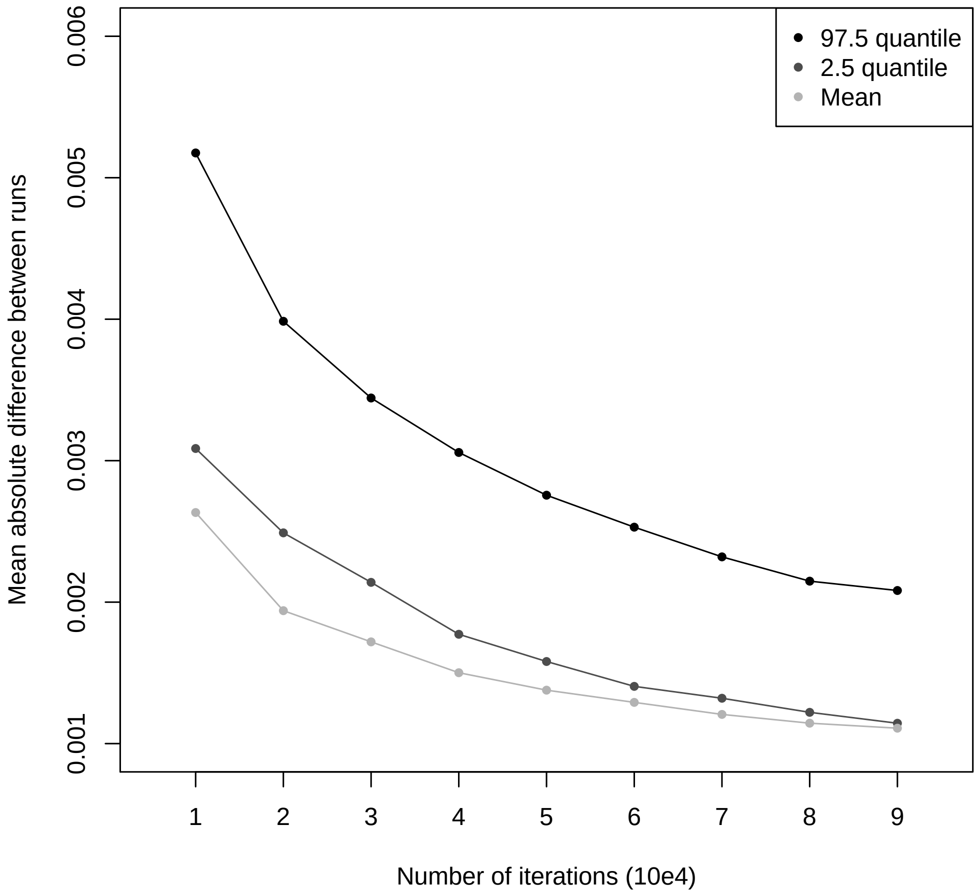
Supplementary Figure 4 compares transcript relative expression estimates $\hat{\pi}_d$ against the simulation truth (see main paper, Section 4.1 for details). Both in Casper-based (left) and Cufflinks-based (right) simulations we observe a higher concentration of points along the diagonal, *i.e.* a stronger correlation between estimates and simulation truth. In particular, Casper with $q_d = 2$ (second row) pushes $\hat{\pi}_d$ away from the boundaries, which helps improve estimation precision.

4. Transcripts missing in current genome annotations

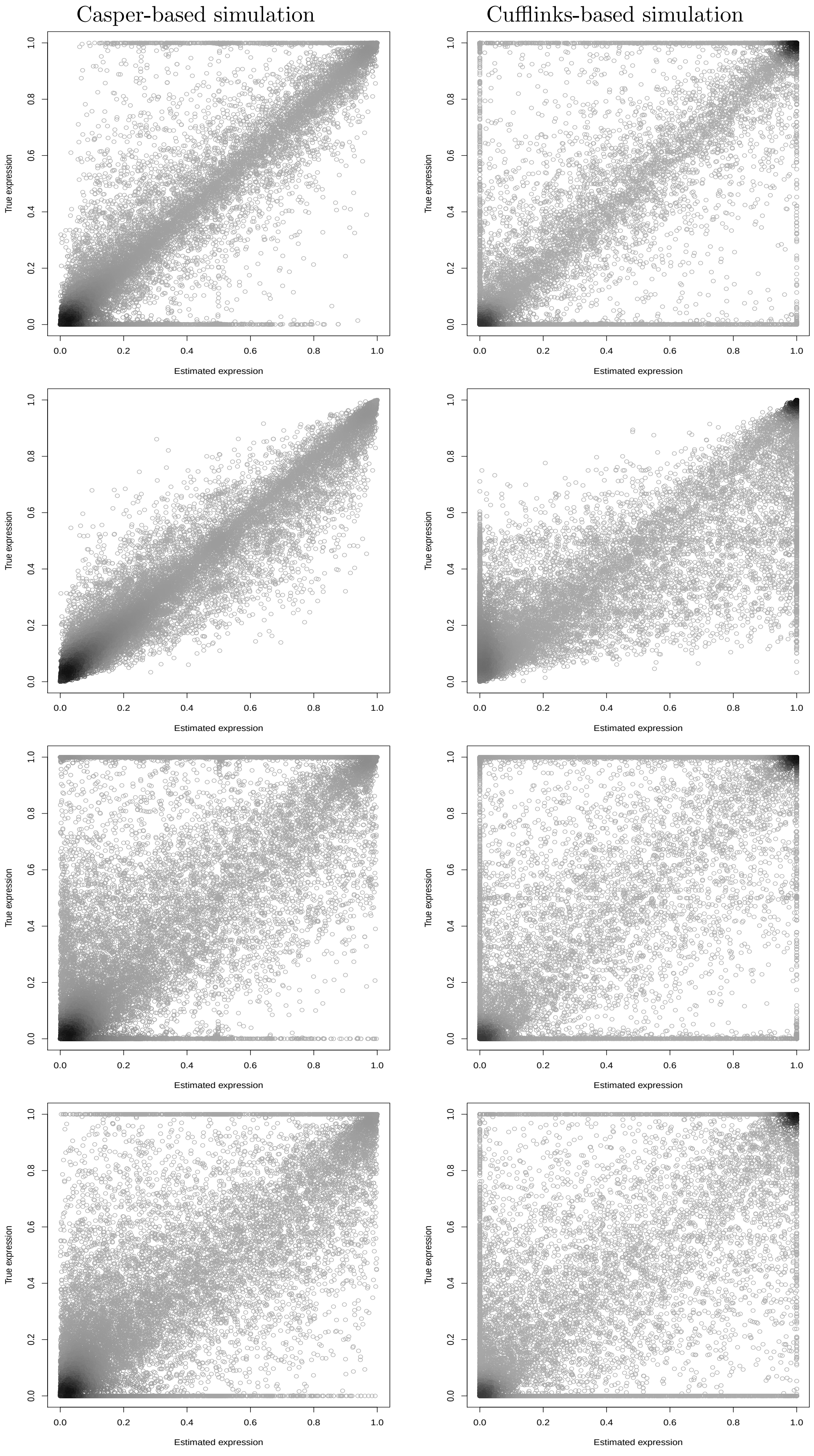
Here we assess the effect of unannotated transcripts on the estimated fragment start and length distributions ($\hat{P}_S$ and $\hat{P}_L$, respectively), as well as on the final estimated transcript abundances.

We used Cufflinks RABT module with default parameters to identify novel transcripts in the two Encode samples (Trapnell et al., 2010; Roberts et al., 2011). Briefly, Cufflinks-RABT adds pseudo-reads generated from the genome annotations to the observed data, and then uses graph theory to predict a parsimonious set of new transcripts to be added to those in the annotations. These new genome annotations were output by Cufflinks-RABT as gtf files and imported into our R package `casper` using function `procGenome`.

In terms of overall expression of the new variants, out of the 40,817,258 read pairs mapping to genes with a single annotated transcript, 33,135,289 were estimated to arise from that transcript (81.1%) and 7,681,969 from new transcripts predicted by Cufflinks-RABT (18.9%). In replicate two 17,529,577 from 22,400,910 reads were assigned to the single known variant (78.2%).

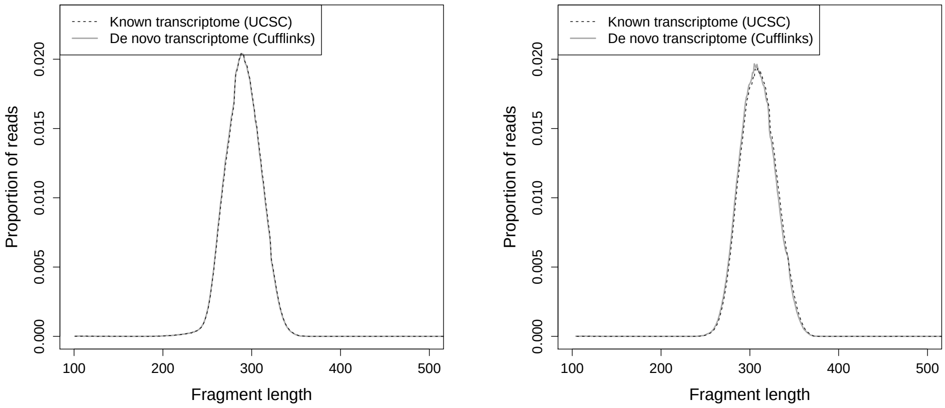


SUPPLEMENTARY FIGURE 3. *Mean absolute difference in posterior mean, 2.5% and 97.5% quantiles between two independent chains (1,000 burn-in) vs. number of iterations*


Casper-based simulation

Cufflinks-based simulation

SUPPLEMENTARY FIGURE 4. *Estimated isoform expression $\hat{\pi}_d$ vs. simulation truth. Row 1: Casper with $q_d = 1$; Row 2: Casper with $q_d = 2$; Row 3: Cufflinks; Row 4: FluxCapacitor*



Supplementary Figure 5. *Encode data. Estimated fragment length distribution P_L considering only UCSC transcripts vs. adding new Cufflinks-RABT predicted transcripts. Left: Replicate 1; Right: Replicate 2*

Further description on the newly found transcripts is provided in Section 4.3 (main paper).

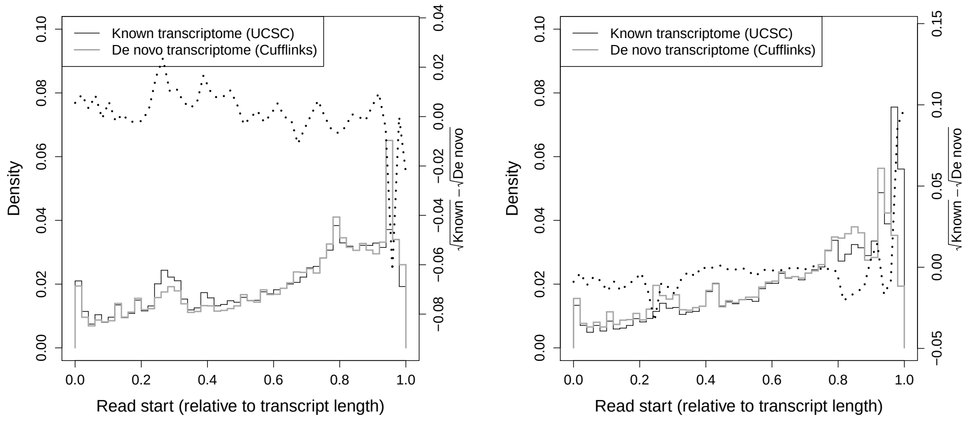
In order to assess the effect of new transcripts, we obtained estimates $(\hat{P}_S^*, \hat{P}_L^*)$ only using genes for which Cufflinks-RABT predicted no new transcripts. In these genes contamination by new transcripts should be minimal. The new estimates were extremely similar to $(\hat{P}_S, \hat{P}_L)$ obtained when using all transcripts (Supplementary Figures 5-6). Next, we assessed the effect on the estimated $\hat{\pi}_d$ induced by using $(\hat{P}_S^*, \hat{P}_L^*)$ instead of $(\hat{P}_S, \hat{P}_L)$. The black points in Supplementary Figure 7 compare $\hat{\pi}_d$ in transcripts for which Cufflinks-RABT found no new transcript, *i.e.* where changes in $\hat{\pi}_d$ are due to $(\hat{P}_S^*, \hat{P}_L^*)$. In most cases the change in $\hat{\pi}_d$ was negligible.

Finally, we assessed the change in $\hat{\pi}_d$ for genes with newly predicted transcripts (Supplementary Figure 7, grey points). As expected, for these genes $\hat{\pi}_d$ decreased so that part of the expression could be assigned to new transcripts.

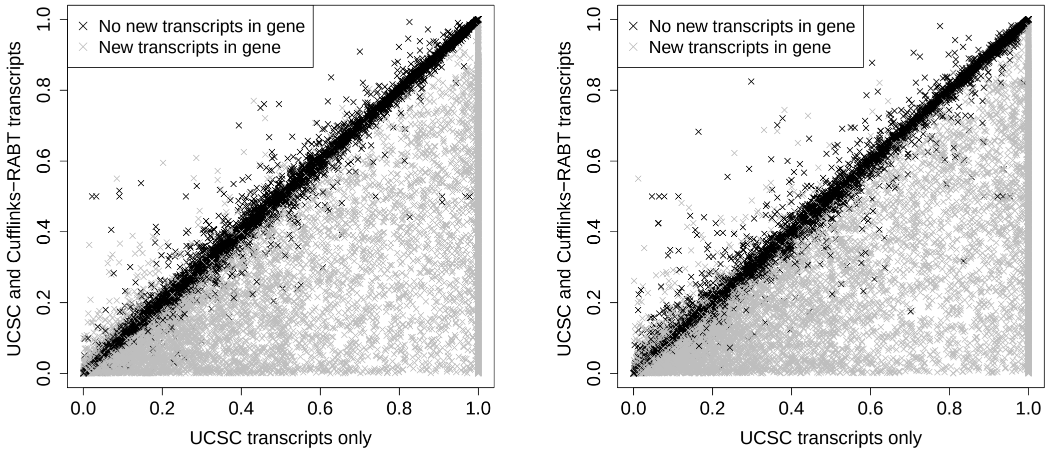
References

Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27** 2325-9.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq



Supplementary Figure 6. *Encode data. Estimated read start distribution P_S considering only UCSC transcripts vs. adding new Cufflinks-RABT predicted transcripts. Black dotted line indicates difference in $\sqrt{P_S}$ (values in secondary y-axis). Left: Replicate 1; Right: Replicate 2*



Supplementary Figure 7. *Encode data. Comparison of $\hat{\pi}_d$ when considering only transcripts in UCSC database vs. adding new transcripts predicted by Cufflinks-RABT. Left: Replicate 1; Right: Replicate 2. Black indicates genes with no new transcripts, grey genes with some new transcripts.*

reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28** 511-5.